Gene-omes built from mRNA-seq not genome DNA

Simple, quick, accurate, less cost, more complete gene sets

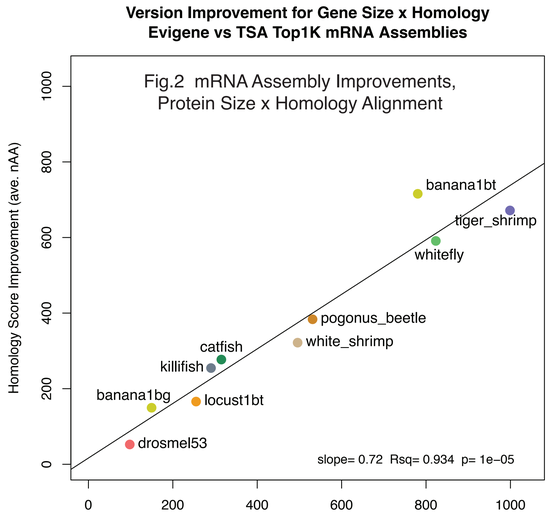
2013 June

EvidentialGene http://arthropods.eugenes.org/EvidentialGene/
Don Gilbert, Biology Dept., Indiana University, Bloomington, IN 47405, gilbertd@indiana.edu

Gene-ome construction with mRNA-seq

For the last 2 decades, complete gene sets have been predicted from gene signal statistics in genomic DNA. The advent of high quality, high volume transcript sequencing provides data suited to constructing genes without statistical guesses, from biological gene products. Informatics methods now have caught up to this data, to construct biologically accurate, measurably complete organism gene sets, or transcriptomes.

Recently improved mRNA assembly methods of the EvidentialGene project (http://arthropods.eugenes.org/EvidentialGene/) are show here with Crustacean, Insect and Tick examples. These methods are relatively simple, rapid and biologically valid; simpler, quicker and better than genome-based predictions. While not yet in general practice, these are recommended as they yeild large improvements to published mRNA assemblies or genome-based predictions. RNA assembly combined with genome-based modelling gives more complete answers, but gene-centric projects will benefit by allocating more effort to transcript sequencing.



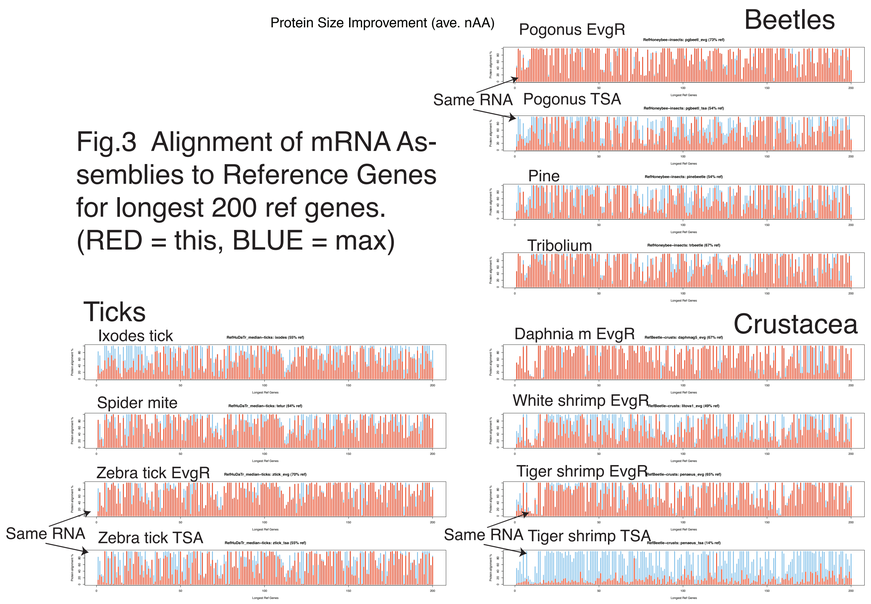
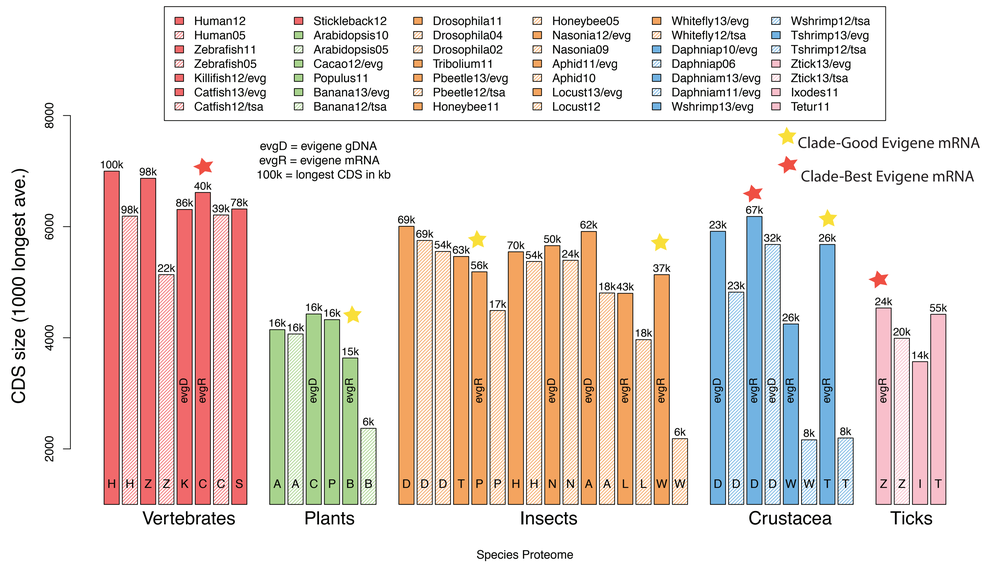


Fig.1 Gene-ome Completeness of Animals & Plants



New or update Gene-omes from EvidentialGene

arthropods/
locust/locust1eg6/ *Locusta migratoria*; pogonusbeetle/pogonus1eg6/ *Pogonus chalceus;*
whitefly/whitefly1eg6/ *Bemisia tabaci*; whiteshrimp/litova1eg6/ *Litopenaeus vannamei;*
tigershrimp/shrimpt1eg6/ *Penaeus monodon;* zebratick/ztick4eg6/ *Rhipicephalus pulchellus;*
plants/banana/banana1eg6/ *Musa acuminata*; cacao/genes/ *Theobroma cacao*;
vertebrates/catfish/catfish1eg6/ *Ictalurus punctatus;*

EvidentialGene mRNA-assembly pipeline

EvidentialGene tr2aacds.pl is my new, somewhat easy to use pipeline for processing large piles of transcript assemblies into a biologically useful "best" set of mRNA, classified into primary and alternate transcripts.

Classification is based primarily on CDS-dna local alignment identity. Transcripts at one locus share exon-sized or larger identities. Perfect fragment CDS are dropped, those with some CDS base differences are kept, with longest CDS as primary transcript. UTR identity is ignored (for now) because many of the mis-assemblies are from joined/mangled genes in UTR region.

Algorithm of tr2aacds:

0. collect input transcripts.tr, produce CDS and AA sequences, work mostly on CDS.
1. perfect redundant removal with **fastanrdb**
2. perfect fragment removal with **cd-hit-est**
3. **blastn**, basic local align high-identity subsequences for alternate tr.
4. classify main/alternate cds, okay & drop subsets by CDS-align, protein metrics.
5. output sequence sets from classifier: okay-main, okay-alts, drops.

See http://eugenes.org/EvidentialGene/about/EvidentialGene_trassembly_pipe.html

Other Evigene scripts for mRNA assembly
evigene/scripts/rnaseq/trformat.pl : regularize and unique IDs in transcript.fasta, adding prefixes for parameter sets.
evigene/prot/namegenes.pl : add gene function names from UniProt and Conserved Domains (CDD) with delta-blastp.
evigene/scripts/rnaseq/**asmrna_trimvec.pl** : process NCBI vector screen, and trim end gaps in transcripts.
evigene/scripts/evgmrna2tsa.pl : check mRNA, add annotation, create public IDs and sequence files, write Genbank TSA format for public submission.

Why mRNA-assembly is better than genome-gene prediction

* Genome assembly gaps, mis-assemblies are common, and disrupt gene models.
* Transposons are a major problem for gene modelling, and abundant.
* Long introns, common w/ transposons, fragment genome-gene models
* Genome gene predictor training on valid species genes is essential and difficult, no training for mRNA assembly.
* Genome gene prediction remains an educated guess. mRNA-assembly is now engineering with reliably good results given accurate data and methods.
* Reference protein mapping needed for genome-genes, but introduces errors. mRNA-assembly without ref proteins can/does produce stronger homology genes with no confounding of evidence.
* mRNA-assembly is simple, quick, accurate, less cost, more complete relative to genome-gene prediction.

- *Daphnia magna* genome-assembly is missing 1/2 expected size to gaps, duplicates. Dmag mRNA genes are now are most complete of crustacea.
- *Ixodes* tick genome genes very fragmented with transposons, Zebra tick mRNA genes much more complete
- Pine beetle genome-gene predictions (Maker) well below *Pogonus* beetle mRNA-genes in homology to reference genes, despite use of ref genes for predictions.

How to build perfect mRNA transcript assemblies
Does & Don'ts

Don't: Use one assembly program, default options. Do: Use several assembly programs and options, as each will provide some better transcripts/genes.

Don't: Make many assemblies then pick one as a best gene set. Do: make millions of transcript assemblies using multiple kmers, programs and other options. Use all assemblies to pick best mRNA per locus. No single assembly is better for all loci than others, because of the large variation in expression levels, gene sizes, types of read errors, etc., that require different options and methods. Use multiple kmer sizes, from read-size down to 25.

Do: Use good RNA de-novo assemblers, even if genome assembly is available. Velvet/Oases, SoapDenovo-Trans, and Trinity produce good assemblies, in that order in my work, and each produces some best assemblies the others miss. Don't: Use Cufflinks only for genome-mapped RNA. Cufflinks underperforms in assembly versus de-novo assemblers, and has more errors of commision (joins) and omission (missing).

Don't: Select longest transcript as best mRNA, as this selects for errors in assembly. Many methods do this implicitly. Do: Select best mRNA with coding-sequence metrics (longest ORF, complete if possible). Longest proteins correlate strongly with strongest homology to other species.

Do: Use at least 200 Million reads of 100bp or longer, mate-paired, to get a complete transcriptome, from current Illimina sequencers. Don't: Use longer 454 reads, due to high error rate. I've not tested very long reads of new machines, but their errors may cause similar problems.

Don't: expect your species/data set to assemble in same way as others have reported. Don't rely on older software without testing newer, and don't expect newer versions to be better (but often they are).

Do: use EvidentialGene scripts and methods for mRNA transcript assembly. The current EvidentialGene_trassembly_pipe is useable by others.
http://eugenes.org/EvidentialGene/about/EvidentialGene_trassembly_pipe.html

RNA assembly can outstrip computing resources. Large memory and cpu clusters are available as shared resources (NSF-XSEDE, others). Digital normalization and genome-mapped partitioning to assemble very large data sets in parts can help.

Genome collaborators and data providers: Daphnia Genome Consortium, Generic Model Organism Database, Indiana Univ. Center for Genomics & Bioinformatics, International Aphid Genomics Consortium, Nasonia jewel wasp Genome project, Cacao chocolate tree Genome project. *Funding*: NSF mostly, NIH, Mars. *Computers*: TeraGrid/XSEDE, NCGAS